BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59157
doi: 10.3897/biss.4.59157

OPEN ACCESS

Conference Abstract

Fast and Easy Access to Central European Biodiversity Data with BIOfid

Christine Driller‡, Markus Koch‡, Giuseppe Abrami§, Wahed Hemati§, Andy Lücking§, Alexander Mehler§, Adrian Pachzelt|, Gerwin Kasperek|

‡ Senckenberg Society for Nature Research, Frankfurt am Main, Germany
§ Text Technology Lab, Goethe-University Frankfurt, Frankfurt am Main, Germany
| University Library Johann Christian Senckenberg, Frankfurt am Main, Germany

Corresponding author: Christine Driller (christine.driller@senckenberg.de)

Received: 30 Sep 2020 | Published: 09 Oct 2020

Citation: Driller C, Koch M, Abrami G, Hemati W, Lücking A, Mehler A, Pachzelt A, Kasperek G (2020) Fast and Easy Access to Central European Biodiversity Data with BIOfid. Biodiversity Information Science and Standards 4: e59157. https://doi.org/10.3897/biss.4.59157

Abstract

The storage of data in public repositories such as the Global Biodiversity Information Facility (GBIF) or the National Center for Biotechnology Information (NCBI) is nowadays stipulated in the policies of many publishers in order to facilitate data replication or proliferation. Species occurrence records contained in legacy printed literature are no exception to this. The extent of their digital and machine-readable availability, however, is still far from matching the existing data volume (Thessen and Parr 2014). But precisely these data are becoming more and more relevant to the investigation of ongoing loss of biodiversity. In order to extract species occurrence records at a larger scale from available publications, one has to apply specialised text mining tools. However, such tools are in short supply especially for scientific literature in the German language.

The Specialised Information Service Biodiversity Research*1 BIOfid (Koch et al. 2017) aims at reducing this desideratum, *inter alia*, by preparing a searchable text corpus semantically enriched by a new kind of multi-label annotation. For this purpose, we feed manual annotations into automatic, machine-learning annotators. This mixture of automatic and manual methods is needed, because BIOfid approaches a new application area with

respect to language (mainly German of the 19th century), text type (biological reports), and linguistic focus (technical and everyday language).

We will present current results of the performance of BIOfid's semantic search engine and the application of independent natural language processing (NLP) tools. Most of these are freely available online, such as *TextImager* (Hemati et al. 2016). We will show how *TextImager* is tied into the BIOfid pipeline and how it is made scalable (e.g. extendible by further modules) and usable on different systems (docker containers).

Further, we will provide a short introduction to generating machine-learning training data using *TextAnnotator* (Abrami et al. 2019) for multi-label annotation. Annotation reproducibility can be assessed by the implementation of inter-annotator agreement methods (Abrami et al. 2020). Beyond taxon recognition and entity linking, we place particular emphasis on location and time information. For this purpose, our annotation tag-set combines general categories and biology-specific categories (including taxonomic names) with location and time ontologies. The application of the annotation categories is regimented by annotation guidelines (Lücking et al. 2020). Within the next years, our work deliverable will be a semantically accessible and data-extractable text corpus of around two million pages. In this way, BIOfid is creating a new valuable resource that expands our knowledge of biodiversity and its determinants.

Keywords

text mining, semantic search, legacy literature, taxon classifier, ontologies

Presenting author

Christine Driller

Presented at

TDWG 2020

Grant title

MO 412/54-1&2, ME 2746/5-1&2, SCHN 1016/46-1&2

References

- Abrami G, Mehler A, Lücking A, Rieb E, Helfrich P (2019) TextAnnotator: A flexible framework for semantic annotations. Proceedings of the Fifteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation. ISA-15, Gothenburg, Sweden.

- Abrami G, Stoeckel M, Mehler A (2020) TextAnnotator: A UIMA Based Tool for the Simultaneous and Collaborative Annotation of Texts. Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France. European Language Resources Association
- Hemati W, Uslu T, Mehler A (2016) TextImager: a Distributed UIMA-based System for NLP. Proceedings of the COLING 2016 System Demonstrations, 2016. Federated Conference on Computer Science and Information Systems, Osaka, Japan.
- Koch M, Kasperek G, Hörnschemeyer T, Mehler A, Weiland C, Hausinger A (2017) Setup of BIOfid, a new Specialised Information Service for Biodiversity Research. Biodiversity Information Science and Standards, 1. https://doi.org/10.3897/tdwgproceedings.1.19803
- Lücking A, Driller C, Abrami G, Pachzelt A, Hemati W, Mehler A (2020) BIOfid annotation guidelines, version 2.8. Goethe University Frankfurt, Text Technology Lab; Senckenberg Nature Research Society; University Library J.C. Senckenberg, Frankfurt am Main.
- Thessen A, Parr CS (2014) Knowledge Extraction and Semantic Annotation of Text from the Encyclopedia of Life . PLOS ONE 9 (3): e89550. https://doi.org/10.1371/journal.pone.0089550

Endnotes

*1 Fachinformationsdienst Biodiversitätsforschung, https://www.biofid.de